DOCUMENT RESUME

ED 230 106 HE 016 167

AUTHOR	Mathews, John
TITLE	The Use of Objective Tests. Teaching in Higher Education Series: 9. Suggestions for the Consideration of Lecturers and Others Concerned with Teaching in Higher Education.
INSTITUTION	Lancaster Univ. (England). School of Education.
REPORT NO	ISBN-0-90169961-6
PUB DATE	80
NOTE	74p.; For related documents, see HE 016 159-166. Produced with the aid of a grant from the University Grants Committee.
PUB TYPE	Guides - Classroom Use - Guides (For Teachers) (052)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	College Instruction; Higher Education; Learning Activities; *Multiple Choice Tests; *Objective Tests; *Student Evaluation; *Test Construction; Test Reliability; Test Use; *Test Validity

ABSTRACT

The use of objective tests in higher education, both as an additional dimension to examinations and as part of the teaching and learning process, is considered. Specific concerns are the characteristics, functions, and limitations and the appropriate systems of organization and management of tests. The following types of objective items are discussed: multiple choice, classification sets, multiple completion, relationship analysis, and true/false. Attention is also directed to the concepts of test reliability (the consistency with which a test measures) and curriculum validity (the degree to which an examination or test matches the curriculum for which it was designed). Next, the practical aspects of objective testing--writing, reviewing, and analysis of objective items--are discussed in detail. Finally, the principles and purposes of objective testing are expanded, and practical guidance is offered for the use of tests in formal examining and as an adjunct to teaching and learning. Consideration is given to specifying quantitatively what the test is intended to measure, the choice and number of items, the arrangement of items, ways that responses can be recorded, and norm- and criterion-referenced tests. Examples of objective test items are appended. (SW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TEACHING IN HIGHER EDUCATION SERIES : 9

ED230106

Suggestions for the consideration of lecturers and others concerned with teaching in higher education

THE USE OF OBJECTIVE TESTS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A. M. Ross

Univ of Lancaster

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

NE 016 167

John Mathews

Teaching in Higher Education Series
ISSN 0309-3565

No.9: The Use of Objective Tests
ISBN 0-90169961-6
© School of Education, University of Lancaster

First published 1977
Reprinted 1978
Reprinted 1980

Published by School of Education, University of Lancaster, Bailrigg, Lancaster, LA1 4YL, England.

Printed in England by Titus Wilson, Kendal.

FOREWORD

This booklet is one of a series produced with the aid of a grant from the University Grants Committee as part of its programme for the development of knowledge about and skill in teaching in higher education. The series is intended to bring together traditional wisdom (not a source to be neglected) research findings when available and relevant, and summaries of what is regarded by the informed as good practice.

The sub-title "Suggestions for the consideration of those concerned ..." echoes that of a famous manual for the guidance of elementary school teachers and this series shares the same purpose. The handbooks may at times be mildly hortatory but they are not meant to be prescriptive. Teaching is something of an art and the aim is to encourage, within sensible limits, something of a personal style. The "rules" (if such they be) are constantly broken by successful teachers but few, if any, get away with breaking all the rules all the time.

Universities and other institutions of higher education have traditionally accepted research and teaching as their principal functions. The systems developed for training in research are well worked out. Till now little has been done for teaching. The initiative taken by the University Grants Committee was widely welcomed and has led to an increasing awareness in higher education generally of the importance of giving to the teaching function the attention it deserves. This series is intended to offer help to those who wish to make an effort to improve their teaching. It is hoped that those who read these pages will, as a result, be better prepared for this important aspect of their daily work.

The authors would be glad to receive reports and illustrative material which readers may wish to send with a view to improving the next edition of these booklets.

C.F. CARTER
Vice-Chancellor
University of Lancaster.

PREFACE

This booklet is intended to be of immediate practical use to teachers in higher education. Its usefulness would be seriously diminished if it lacked specific examples; the author is indebted to those who have generously permitted him to reproduce items taken from objective tests which have been used in higher education or prepared by expert and experienced test writers. He would like to thank

The Associated Examining Board,
The Open University,
The Educational Division of the Chemical Society,
The Department of Biological Sciences, University of Lancaster, and
the authors and publishers of Multiple Choice Examinations in Medicine by J.P. Hubbard & W.V. Clemens, published in Philadelphia by Lea & Febiger.

CONTENTS

INTRODUCTION ---------- i

PART I: HOW OBJECTIVE ARE OBJECTIVE TESTS? ---------- 1

1. Objectivity and fixed response ---------- 1
2. Types of objective item ---------- 4
3. Objective tests and curriculum validity ---------- 6
4. Reliability ---------- 12

PART II: TESTING IN PRACTICE ---------- 15

1. Writing objective items: the stem, the responses, editing ---------- 15
2. Reviewing objective items: a checklist ---------- 20
3. Item analysis: facility, discrimination, favoured wrong response ---------- 22
4. Pre-tests and item banks ---------- 28

PART III: THE USE OF OBJECTIVE TESTS ---------- 31

1. Objective tests in examinations ---------- 31
2. Objective tests in teaching and learning ---------- 43

PART IV: IN CONCLUSION ---------- 51

APPENDIX: Examples of objective items.

INTRODUCTION

Teachers in higher education may have noticed that an increasing number of their students have experienced objective tests in schools; since the mid-1960s the public examination boards have introduced these tests at all levels and in many subjects. This, in itself, is not sufficient reason for introducing such tests into higher education; but the increase in objective testing at school level is not mere fashion, although doubtless a desire to appear up-to-date has played a part. There are strong educational reasons for the use of these tests of which teachers in higher education could take note with benefit to themselves and to their students. It may be argued that teaching and learning in higher education is substantially different from that in schools and requires, therefore, different techniques of assessment. Strangely, this argument is not used against the ubiquitous essay. And, surely, to anyone who has experience of both sixth form and undergraduate teaching the argument is specious; neither the teaching nor the students change so much over one summer vacation.

It is not, however, the author's intention to "sell" objective tests to teachers in higher education nor to engage in a sterile argument on the relative merits of objective tests and the traditional essay. In any event, objective tests are not new to higher education; they have been used for many years by the medical schools;

some time ago the Chemical Society established an objective item bank which is used by colleges, polytechnics, and universities. The Open University in its Computer Marked Assignment system has attained a peak of complexity in the design of objective tests and in the subsequent analysis of them, and doubtless there are many other instances of the use of these tests. But, still, in many sections of higher education objective tests are used not at all and it is to further the principle of diversity, advocated in the booklet "Inside Assessment", that this introduction to objective testing is written.

Those who are experienced in the design and use of objective tests will be familiar with much that is written here, although there may be some sections, particularly that on the use of the tests as an aid to teaching, in which they will find something of interest. The readers towards whom this booklet is mainly directed are those who have not yet embarked on objective tests, yet are prepared to consider the various issues involved in their use in higher education, both as an additional dimension to examinations and as part of the teaching and learning process. To this end the characteristics, the functions, the limitations and the appropriate systems of organization and management of the tests, are set out so that those teachers in higher education may be helped to decide objectively whether to use them or not.

The first part of this booklet, "Part I: How Objective are Objective Tests?" discusses the nature of this mode of assessment, the purposes for which it may be used in higher education, the types of test in common use and the principle of validity, here defined as curriculum

validity, which the examiner should observe if his tests are to be fair in practice.

"Part II: Testing in Practice" describes the practical business of objective testing - the writing, reviewing and analysis of objective items - in sufficient detail for this to be a manual for those who are new to objective testing.

"Part III: The Use of Objective Tests" extends the discussion of the principles and purposes of objective testing and gives additional practical guidance for the use of tests in formal examining and as an adjunct to teaching and learning. The author's own view of the value of objective tests for the teacher and examiner in higher education is set out in "Part IV: In Conclusion".

PART I

HOW OBJECTIVE ARE OBJECTIVE TESTS?

1. OBJECTIVITY AND FIXED RESPONSE

If by objective tests is meant a form of measurement which is not subject to the decisions of individuals, they are objective in one respect only, the marking of them. Even this is not entirely true: human error can still arise, if rarely, in the marking of an objective test. The name is an unfortunate one, implying an ideal impartiality and independence of human whim which is often far from the reality. In their construction, their specifications, and in the writing of the individual questions, objective tests are as likely to be influenced by the judgment and opinion of individuals as most other forms of assessment; they are, then, far from objective.

The distinguishing characteristic of an objective test item* is not so much its objectivity as the fact that the correct response to it is predetermined by its author(s). For this reason the description fixed response has been

---

* *"Item" is test jargon for what is usually called a question. However, since many items are not actually questions, there is some logic behind the use of the term item and it is used throughout this booklet.*

applied[1] to this type of test item: a student either answers exactly as the examiner wishes and gains a mark, or he does not and gains zero or even a penalty mark.

It follows that if a student wishes to be successful in an objective test he must think as the examiner thinks; there is no room for divergence. In an examination situation this is the determining factor. If classes of degree are in part determined by the result, it would be a strong-minded student indeed who would give answers with which he feels the examiner may disagree; free response is unlikely in such circumstances. Let us not suppose, however, that this is uniquely characteristic of objective tests. Many "essay" questions, while apparently offering an invitation to divergence, discussion, and freedom to disagree with the examiner, hide a rigid marking scheme which is anything but free. At least the objective test has the virtue of honesty in that the examiner has the courage to write down in the question itself what he thinks is the best answer. This convergent aspect of objective tests applies less when they are used in teaching situations where the total mark which a student attains matters less or not at all. These two different aspects of objective tests will be discussed in more detail in later sections.

The fixed response requirement is, perhaps, the only truly essential characteristic of all the many kinds of objective test item but, for the purposes of this booklet we add

---

(1) SCHOOLS COUNCIL (1973) Examinations Bulletin 27. *Assessment of attainment in sixth form science.* London. Evans/Methuen.

another defining characteristic: an objective test is one in which the student is required to select his response from a set of responses all written by the examiner and only one of which is normally accepted. We exclude, therefore, the form of open-ended question which requires a single word or phrase by way of answer although that response may be equally predetermined by the examiner and the question may be equally objective.

It is, however, possible to construct items in a variety of different and sometimes complex forms. Items 30-41 in the Appendix are examples taken from the Open University Computer Marked Assignment in which 12 responses are required from a possible total of 25. Items like this give rise to many permutations of responses and presumably each carries various proportions of the maximum marks allocated to the item. This contrasts with simple items in which there is only one correct response normally carrying only one mark - Item 1 in the Appendix serves as an example. But whatever its degree of complexity an item retains its two essential characteristics: the responses which will carry marks are predetermined by the examiner and the responses are printed as part of the question.

In passing it should be said that complex items of the type illustrated by Items 30-41 are usually the product of experienced construction teams. While many similar and very interesting examples can be found in the computer marked assessment system of the Open University, those attempting to design objective tests for the first time, to whom this booklet is addressed, would be wise to start with the simple ones.

2. TYPES OF OBJECTIVE ITEM

There are many kinds of item with the two necessary characteristics, a predetermined correct response and a set of responses from which one is chosen. Several will be described br.efly but those who are new to objective testing are advised to confine themselves at first to the multiple choice form. Examples of this and of other forms are given in the Appendix; it is in conjunction with the Appendix that this section should be read.

Multiple Choice

The structure of a multiple choice item is in two parts: the stem in which the question is posed and necessary information can be given, and the responses, usually four or five in number. Sometimes the incorrect responses are called distractors. This structure will be elaborated in the section on item writing (p.15). A set of multiple choice items is sometimes constructed so that all are dependent on the same information but they are not usually dependent on each other. Items 5-8 in the Appendix provide an example of such a set.

Classification Sets

Sometimes called matching pairs, classification sets require students to match one set of entities against another. There are many variations; a common form, often used in science, is to list four or five general categories; the students are then required to classify specific examples. Closer examination of this type of item reveals that it is simply a special case of the multiple choice; the format, however, allows a larger

number of questions to be condensed into less space.

Multiple Completion

This type of item allows a choice of a single combination of responses rather than a single response. There are many variations of this type of item, the examples in the Appendix being only some among many. In large scale examining, particularly when texts are mechanically marked, it is usual to award one mark for the choice of the correct combination and no marks for any other combination even if it is partly correct. This simple system of scoring is used largely for technical convenience and there is nothing to prevent the use of a more subtle scoring system, particularly if the test is being used by a teacher for diagnostic purposes. The Open University has adopted a flexible scoring system in many of the items in its computer marked assignments.

Those who are interested in expanding their use of objective testing, and have access to computer facilities, should note that the special type of multiple completion item of the form n from 5, n denoting more than one correct response, is increasingly used in higher education. Examples are items 42-43 in the Appendix.

Relationship Analysis

Items of this type rarely live up to their rather fine title. They consist of two statements, the first usually called an assertion, the second a reason. Students have first to decide whether each, considered separately, is true or false; then, if both are true, whether the reason is a correct explanation of the assertion. At best they test the ability of a student to link the truth of a

specific observation with a more conceptual statement which purports to explain that observation. This can only occur when both statements, separately, are true. At worst the items are no more than a variant of the true/false type, but what constitutes truth and falsity may often be disputed. Items like this may be of more use in some subjects than in others; the reader must judge for himself after looking at the examples in the Appendix.

True/False

It is now rare to find this, the simplest of all types of objective item, used in large scale examining. The effects of random guessing can be high unless a very large number of items is used. However they are easier to write than many others and they can be answered quickly. Their use as a teaching aid for diagnostic purposes, particularly in schools, should not be overlooked. Whether they have general application to teaching in higher education, where niceties of distinction often blur the stark true/false issue, remains to be seen.

3. OBJECTIVE TESTS AND CURRICULUM VALIDITY

Examiners are constantly urged to consider the validity of their examinations. Yet validity is a concept used nowadays so loosely as to have become almost meaningless. In the sphere of examinations and tests the most hackneyed definition is: the effectiveness of an examination or test in measuring what it is intended to measure. (The assumption that examiners always know what it is they intend to measure is by no means valid.) A valid test,

to quote the Shorter Oxford English Dictionary, would be one that was "well founded and applicable; sound and to the point." The question remains, applicable to what, and to which point.

The view taken in this booklet is that an examination should be well founded in and applicable to the teaching and learning which it serves, in short the curriculum. Hence the concept of curriculum validity: the degree to which an examination or test matches the curriculum for which it was designed. To put it another way: the structure and specification of an examination should match the structure and specification of the curriculum. Or yet another way: the situation in which a student finds himself in an examination should not be significantly different from that which he has experienced in his learning situation.

Difficulties arise, of course, if the curriculum has no structure or specification or if the learning situation is diffuse and ill-defined. But rarely is a curriculum completely spontaneous; analysis usually reveals that some design is there, even if it is implicit rather than explicit. A crude analysis of most curricula gives rise to three main dimensions:

* subject matter
* expected student knowledge, skills and abilities,
* teaching/learning activities.

The detailed specification of a curriculum in respect of those dimensions is usually thought to be easier to achieve in the sciences, including the applied and social sciences, mathematics and business studies, than it is in

the arts and humanities. It should not too readily be assumed that this is the case and that courses in the arts and humanities are too nebulous to be analysed in this way. Such an analysis is, of course, the first step in deciding whether objective tests can make a contribution to the assessment of attainment in any course.

It is not always possible nor desirable to quantify these dimensions but recognition of them, if only qualitatively, is the beginning of all attempts to estimate the degree to which they are reflected in assessment procedures of any kind and hence their validity. Such an estimate may be called curriculum validity.

At its simplest level, any attempt to ensure that the assessment situation bears resemblance to the teaching/learning situation is a move towards improving the curriculum validity of the assessment. If, for example, teaching and learning are mainly concerned with discussions of problems to which there are no definitive solutions, a test which sought only the recall of factual information would have low validity. If a course was mainly concerned with the development of practical skills, it is unlikely that assessment in the form of written essays would have high validity. Nor is an objective test likely to be appropriate for an oral language course.

It is most unlikely that objective tests alone could satisfactorily meet the requirement of curriculum validity for examinations in higher education. The same could be said with equal truth of essay tests or, for that matter, any other single assessment instrument. It is necessary, therefore, to set out the distinctive contributions which

objective tests can make to the curriculum validity of a balanced examination system. Each of the three dimensions listed above will be considered in turn.

(i) Subject Matter

A distinctive feature of objective tests is the large number of questions which can be answered in a short period of time; 40 items per hour is about the norm and a higher rate is not uncommon. It follows that the range of subject matter which can be tested is much greater than that permitted by the more traditional forms of examination which typically allow one or two questions per hour.

Of course a topic cannot be explored to the same depth in a single objective item as it may be in a single essay or similar type of question; and it could be that objective items test attributes of students which are different from those tested by essay questions. It may be argued, also, that it does not matter if students are tested on only a fraction of the material which they have studied during a course. This latter argument is difficult to defend; students may ask why so much of the material they were required to study can be declared redundant when the time of assessment arrives. If they are allowed to specialize in the examination, why were they not allowed to specialize in their studies? They should at least be informed of the areas of choice so that the elements of chance and of question spotting in the traditional essay examination are reduced.

In objective tests the correlation between curriculum subject matter and examination subject matter can be

very much higher. There is, however, an important qualification: since objective tests do not allow a choice of questions, it is necessary to ensure that the subject matter tested is that which is commonly studied by all students. A feature of recent trends in examining is a clearer distinction between material which is part of a core curriculum and material within which some freedom of choice is allowed. It is important to recognize that objective tests relate to the former, except possibly when they are used diagnostically as a teaching aid.

(ii) Student Attributes

This term is used loosely to cover those skills, abilities and attitudes which a course is thought likely to foster. There are many ways in which these attributes may be classified - some of the best-known and most frequently used classifications are described in "Inside Assessment". Some discussion is necessary here about which attributes can be most appropriately assessed by the use of objective tests.

A cursory analysis of objective items often reveals that skill in remembering predominates; and there is no doubt that they can test memory very economically in terms of items per unit time. (In passing, it should be mentioned that the act of remembering in an objective test is one of recognition rather than of open recall.) But it would be wrong to assume that this is their only function, or that they are most effective when testing memory. There are now many instances of the use of objective items to test higher cognitive skills. Although they are not easy to construct, items testing ability to analyse

information, to solve problems, and to make judgments, are not uncommon; frequently they show objective testing at its best. The reader is invited to find some amongst thore quoted in the Appendix. So it would be wrong to suppose that objective tests cannot contribute to curriculum validity when applied to courses which demand higher cognitive skills than that of recalling. On the other hand one should view tests which purport to assess divergence, creativity, and originality, with some suspicion. Personality traits pointing towards divergence may well be assessed to some degree by carefully constructed tests; but to suppose that the choosing of one of four or five predetermined responses can be a creative act is to make nonsense of the term. It is interesting to note that one group of items written for the Open University, and quoted in the Appendix, contains the instruction, "Remember that your answer should be based on the views contained in Unit 1 and NOT on your personal views". In other words, the items are designed to test comprehension of a text and thought convergent with that of the examiner, not divergent or non-conformist thought about the statements in the items.

(iii) Teaching/Learning Activities

The activities in which a student engages when answering an objective test amount to reading, thinking and making a mark with a pencil. It would be idle to suppose that the tests encourage oral, manipulative, or discursive skills. Nor could the tests be said to resemble the activities of construction, composition or artistic creation. This is not to say that it is not possible to construct tests the subject matter of which concerns these activities; tests about technical skill and artistic

composition are available and have been used. The point is that the actual activity in the test situation can be remote from that of the learning situation and inasmuch has low curriculum validity.

Of course, the same questions could be put to those who require in their examinations no more than the writing of essays. The writing of essays has little to do with what happens in real life; and some would say little to do with what happens in real education. A good objective item can involve the analysis of evidence and the choice of alternative solutions to a problem all of which have some merit; this may be closer to real-life activities than the rather impersonal form of the test item might lead us to suppose.

In summary, the limitation of the contribution of objective tests to curriculum validity has been admitted. But their contribution is no less, and may well be greater, than that of any other single assessment technique. The trend in assessment in schools is to use a variety of techniques, each making a particular contribution to validity. There is every reason to suppose that a similar blend would be equally effective in higher education.

4. RELIABILITY

It is not sufficient simply to specify what it is that a test is intended to measure; it is equally important to give some consideration to the precision with which the measurement is made. The concept of validity,

therefore, cannot be divorced from that of reliability which may be defined as the consistency with which a test measures. In general it can be claimed that well-constructed objective tests have high reliability compared with traditional forms of examination.

This is not the place for a long discourse on the various methods by which the reliability of tests can be estimated; sufficient to say that all measures of reliability are estimates and should be interpreted with caution. Ideally, the scores of a group of candidates in a test should be compared with the scores on the same test after an interval of time and assuming i) that no additional learning or diminution of learning had taken place and ii) that memory of the test items themselves had been erased. The degree of correlation between the two sets of scores would be a measure of the reliability of the test. Clearly, the fulfilment of such conditions is not possible in practice and it is necessary to resort to various artifices in attempts to simulate the conditions. The mathematical outcome of these devices can be found in the many texts on the statistics of testing,[1] particularly in the form of the Spearman-Brown and Kuder-Richardson formulae. Whether it is worthwhile to go to the trouble of calculating the reliability depends on the extent to which the test is to be used as a discriminating instrument in an examination and on the statistical leanings of the examiner. In large scale examining at school level, supported by computerized processing of the scores, calculation of the reliability of objective tests is standard practice. But

---

(1) LEWIS, D.G. (1967) *Statistical Methods in Education*. London. University of London Press.

if the tests are used solely as diagnostic instruments to assist teaching, calculations of their reliability are unnecessary.

In passing, it must be mentioned that the consistency with which a test makes measurements does not, in itself, ensure that it is accurately measuring what it was intended to measure. It is possible for tests to measure consistently attributes of students which differ completely from those which the examiner intended to measure. It follows that a test may have high reliability but low validity. It is not possible, however, for a test to have high validity if it has low reliability; validity is a function not only of <u>what</u> is measured but of <u>how accurately</u> it is measured.

PART II

TESTING IN PRACTICE

1. WRITING OBJECTIVE ITEMS

The ability to construct good objective items requires a blend of art, skill and experience; and there are some who never acquire it. To them one's advice must be to pursue other forms of testing, because there is nothing quite so bad as a bad objective item. Those who persevere should take heed of the first premise: objective item writing is best achieved as a group activity. There are few, indeed, who can consistently write good items without the assistance of colleagues. Of course the first idea and draft of an item is necessarily the work of an individual, but thereafter it is uniform practice to submit the items to review by a group of three or four people who diligently seek to find fault in them.

Most teachers, given a little practice, will find they can write items which are more than adequate for most purposes; it is not so difficult a task. Write a few only at a time; then wait until inspiration returns. While not quite on the level of writing poetry, objective item writing is a creative art which, if well done, can be tiring.

The following guidelines for writers relate principally to multiple choice items but much of what is written can be applied to any of the other objective forms. It is useful at first to consider separately three parts to each item:

* the stem, in which the question is posed and any necessary information given,
* the correct response, that is the writer's answer to the question, and
* the distractors or incorrect responses.

The stem

The easiest type of stem to write is that which simply asks for the recall of information. In the Appendix, Item 4 is an example; although it is phrased in the form of an incomplete statement, it could equally well have been phrased as a direct question. Provided that the question is a clear and unambiguous demand for information, and provided that it gives the appropriate cue for the recognition of the correct response, the construction of such an item is fairly straightforward.

If, however, one is seeking to test more complex skills such as those of analysis, application and judgment, rather than that of recall, it is likely that a good deal more effort needs to be put into the construction of the stem. It may then be more profitable to make the stem considerably longer, putting into it information which the student has to analyse and manipulate in a manner much more demanding than simple recall or recognition of information. The information can be varied almost infinitely; the stem of the set of items 5-8 in the Appendix provides an example. Other information could

be diagrammatic, graphical, tabular, photographic; anything which provides material on which the student is required to work rather than acting as a cue for recall.

The information-giving type of stem is not free from problems, of course. There is a good deal more material for the student to read, slowing down the rate of progress through the test and penalizing those with reading difficulties. It also exposes the competence or incompetence of the writer; the more he writes the more likely is he to get his material wrong and exposure to criticism becomes greater. The point about speed of reading can be a real difficulty which should be kept in mind when reviewing the test as a whole. But criticism of item writers is a healthy trend; critical analysis of objective items is part of the game; for too long examiners have had it too easy.

All the information given must, of course, be essential information and not padding. It should be self-evident that the language used both in presenting information and in framing questions must be precise and economical. Many questions have to be read rapidly in a short time. This is in itself a confusing activity; students should be spared the further hazard of vague phrasing and complicating clauses. Negative questions should be used sparingly, double negatives avoided and direct questions preferred to indirect or implied questions.

Ideas for objective items are not always easy to come by; good ideas should never be lost. An element of originality is often needed for the questions that test higher skills. Good questions reflect good teaching and ideas

are more likely to come to mind while actively teaching or reading than while determinedly seeking inspiration - make a note and develop it later.

Finally make sure that the stem leaves the student in no doubt about what he is to do.

The responses

The same kind of advice applies to the writing of responses as was given for the stem. The best responses, both correct and incorrect, are those which a teacher would expect from a group of students if the question had been put in an open-ended way. In other words the distractors should reflect errors which experience has shown to arise in teaching situations. Above all, try to read them through your students' eyes.

The correct response does not necessarily have to be an absolute truth - a rare commodity in any event; students can be asked to select the best answer from among four or five others of varying degrees of plausibility. The phrase "which is the most likely" is one frequently found in the stem of objective items; Item 5 in the Appendix provides an example. Those first starting to write objective items, however, usually prefer to ask for a response which is clearly "correct" rather than "most likely". In this event they should make sure that it is actually correct; it is remarkable how many questioners fail to answer their own questions.

Since the incorrect responses must have some plausibility if they are not to be redundant, they can be used to make the question harder or easier. The nearer they are in

meaning to the correct response and to each other the more difficult will a student find it to choose correctly. It is for this reason that multiple choice questions can be more difficult than those asked in an open-ended form; although the correct answer is there to be recognized rather than recalled, the distraction of the other closely related responses makes decision less easy. The converse also applies, of course; and distractors which are clearly irrelevant to the question can lead students to the correct response by mere elimination of the others.

Editing

On the large scale objective tests are used mainly by national examination boards who edit their tests stringently. A consistent house style is most rigorously applied - giving at times the impression of pedantry - but this is done to maintain precision of language as well as consistency of style and form. It is surprising how small a defect in an item can deflect at least some of the candidates to or away from the correct response. In small scale testing for teaching purposes limitations of time and resources mean it will not often be possible to edit items so thoroughly.

Nevertheless the solitary writer should make a habit of meticulously going over his items himself before submitting them to the systematic reviewing procedure described in the next section; he should try to develop a good consistent style and form of his own. In matters of style, form, phrasing, terminology and punctuation he should establish strict rules. Some will be self-evident, some will reflect his own tastes, most should derive from the prescriptions of those with skill and experience in objective testing. The watchwords should be precise, concise, unambiguous,

consistent, and, where possible, grammatical.

2. REVIEWING OBJECTIVE ITEMS

Some commodities improve with keeping, whisky and - for a much shorter period - curry, come to mind. Others, amongst which objective items must be numbered, often quickly lose whatever charm they had. Perhaps the most important advice an item writer is ever given is to set every item aside for a day or two to be read with fresh eyes before it is included in a test.

A rigorous review (called shredding by some) of objective items is a necessary part of test construction. In large scale examining they will be reviewed on more than one occasion and usually by a group of people; this is the ideal. On a small scale it may be that the best you can hope for is a review by a colleague. Failing all else, at least check the items yourself using the procedure set out below. Far better to find the faults yourself before the event than to leave your students to tell you about them after it.

A CHECKLIST FOR ITEM REVIEW

1. Read and answer the question
   - 1.1 Choose your response before looking at the key.
   - 1.2 If you get it wrong, which is at fault - you or the question?
   - 1.3 Is there more than one correct response?
   - 1.4 Is there an ambiguity in the question?
   - 1.5 Is there an ambiguity in the response(s)?
   - 1.6 Can it be answered in the brief time available?

2. Check the stem

2.1 Is the question important enough to be asked?

2.2 Is there sufficient information to answer it?

2.3 Is it dependent on other items in the test?

2.4 Is it precise, concise and grammatically correct?

2.5 Is it a positive question, a negative question, or a completion?

2.6 If a negative, is the negative highlighted?

2.7 Does the question ask about a universal or a particular truth? (There are very few of the former.)

3. Check the responses

3.1 Have all got some plausibility?

3.2 Does the correct response contain a superficial clue? (Perhaps a word repeated from the stem.)

3.3 Does a double negative arise, one in the response and one in the stem? (Can be confusing.)

3.4 Are they all about the same length?

3.5 Are they visually confusing? (Often arises when mathematical formulae are used.)

3.6 If they are completions of the stem, do they all complete grammatically?

3.7 Do they all contain a repeating phrase which could best be written once in the stem?

4. Check the format

4.1 Same number of responses in each item. (Not absolutely necessary - see some of the Open University Items - but is normal practice.)

4.2 Consistent lettering and numbering.

4.3 Each response on a separate line unless all fit onto one line.

4.4. Directions, stems and responses all on the same page.

4.5 Numerical responses sequential, usually the smallest first.

5. Classify the item

5.1 What kind of student attribute does it test?

5.2 What area of subject matter does it test?

5.3 Estimate its level of difficulty (easy, difficult or average will do).

6. Judge the item

6.1 Acceptable as it stands.

6.2 Acceptable with some immediate small alterations.

6.3 Return to writer for reconstruction.

6.4 Waste paper.

3. ITEM ANALYSIS

The review procedure set out above is subjective; it is used to analyse and evaluate items before they are used. But objective items have an advantage over most other forms of assessment; their performance under operational conditions can be quantitatively analysed and evaluated after they have been used. Item analysis is a technical term used to describe measurements made on tests after they have been used.

Item analysis yields information of three main kinds. For each item a descriptive measure can be obtained of

* facility (or difficulty)
* discriminating power
* which wrong responses were most favoured.

The simple methods used for calculating each of these statistics will be described in turn.

Facility (F)

This is a simple statistic which shows how easy or difficult an item has proved to be. It is usually shown either as a percentage or a decimal fraction. For example:-

Total number of candidates (N)	= 90
Number of correct responses to Item (R)	= 63
Facility of Item (F) = $\frac{R}{N}$	= $\frac{63}{90}$
	= 0.70 or 70%.

Sometimes a refinement can be introduced by taking account of the number of candidates who did not attempt that item but this number is usually very small and can be disregarded for most purposes.

Discrimination (D)

This statistic shows the degree to which a particular item discriminates between the able and less able students. Ability is here defined in terms of how well students do on the test as a whole. If, then, a particular item is to contribute to the discriminatory function of the whole test, the more able students should obtain a greater proportion of correct responses than do the less able. There are several ways in which this can be quantified: in computerized analysis the biserial or point-biserial coefficients are used. However, since we are concerned here mainly with the use of objective tests on a small scale, the following simpler procedure will be described.

1. Arrange the students in rank order according to their score on the test as a whole.
2. Separate the answer sheets of the top third and the bottom third* of students.
3. For each item count the number of correct responses (N_H) obtained by the top third and count the number of correct responses (N_L) obtained by the bottom third.
4. Count the number (n) of students constituting one third of all the students who took the test.
5. Calculate the discrimination of the item (D).

$$D = \frac{N_H - N_L}{n}$$

For example:

n = 30 (one third of all students)

N_H = 27 i.e. 27 out of 30 answered correctly

N_L = 15 i.e. 15 out of 30 answered correctly

$$D = \frac{27 - 15}{30} = \frac{12}{30} = \underline{\underline{0.40}}$$

It follows that if an item is to contribute to the total discriminatory power of the test, N_H must be greater than N_L and, therefore, D must be positive. The two extremes would be:

i) $n = 30$, $N_H = 30$, $N_L = 0$

therefore $D = \frac{30 - 0}{30} = 1$

ii) $n = 30$, $N_H = 0$, $N_L = 30$

therefore $D = \frac{0 - 30}{30} = -1$

---

* *The third is not rigid; it could be 30% or a quarter or a half.*

And in the middle:

iii) $n = 30$, $N_H = 20$, $N_L = 20$

therefore $D = \frac{20 - 20}{30} = 0$.

In (i) the item is discriminating totally between the able and the less able students and is making maximum contribution to the total discriminating power of the test. In (ii) the item is also discriminating totally but in the opposite direction; that is the able students on the test as a whole are answering incorrectly and vice versa. This item, therefore, is decreasing the discriminatory power of the test as a whole. In (iii) the item is making no distinction between the able and less able candidates; it is having neither a positive nor a negative effect on the discriminating power of the whole test.

Opinion varies on what constitutes acceptable limits to Facility and Discrimination Indices. Facility Indices between 30 and 80 could be accepted for items of the four response type, although some authorities suggest narrower limits. For Discrimination Indices the lower limit is the only one to consider since it is usual to assume that the higher the Discrimination Index the better. Again, there are no firm rules; normally a positive value for D is considered essential, but the lower limit depends on the formula used to calculate D and the subjective judgment of those constructing the test. In the author's experience, test constructors are often willing to accept items with a lower D value than most of the books on objective testing suggest, particularly if items are in short supply.

The various decisions concerned with Facility and Discrimination Indices depend largely on the purpose of the test as a whole. Much of the literature on testing relates to the mass examining of large numbers of candidates and it may be that the limits of what is recognised as statistically proper may be narrower than in other circumstances. The bias of this booklet on objective tests is towards their use with small numbers of candidates; and it is concerned just as much with the teaching and diagnostic function of the tests as with the discriminatory function. Once the discriminatory function of the tests becomes secondary, one can view the F and D statistics quite differently. This point will be taken up in more detail in the section dealing with the use of objective tests as a teaching aid.

Favoured Wrong Responses

This, the third statistic, shows which distractors (incorrect responses) attracted the greatest proportion of the students. It is often best set out in the form of a simple matrix showing how many students chose each response (* indicates the correct response).

	A	B*	C	D	Total
Top third	2	27	1	0	30
Middle third	6	21	3	0	30
Bottom third	10	15	5	0	30
Total	18	63	9	0	90

F = 0.70, D = 0.40.

The foregoing example shows that A is the favoured wrong response and since it attracted a greater proportion of the top students compared with the bottom students it is unlikely to give cause for concern. But there would be cause for concern if a matrix was as follows:-

	A	B*	C	D	Total
Top third	20	8	0	2	30
Middle third	15	12	0	3	30
Bottom third	10	16	0	4	30
Total	45	36	0	9	90

F = 0.40, D = -0.27.

This example shows first of all a negative Discrimination: more of the bottom students chose the correct response that did the top students. Second, an apparently incorrect response, A, attracted the greatest proportion of the top students. An analysis of this sort demands that the item be reviewed again to try to find out what it was about the item as a whole and response A in particular which gave rise to this anomalous result.

In addition, it will be seen that in both examples there was one response which attracted no choice at all, response D in the first example and response C in the second. Such an occurrence is called a non-functioning response and it signifies that the item might just as well have had only three responses as the prescribed four. If the item is to be used as a teaching aid this does not much matter;

but, if it is to be used in an examination it is necessary that the non-functioning response be re-written to make it more plausible.

4. PRE-TESTS AND ITEM BANKS

In large scale examining it is common practice to try out the objective items on a sample population of candidates. This takes the form of two or more pilot tests containing items which have been through the reviewing process, the aim being to determine the Facility and Discrimination Indices and to reveal any defects which the items may have. The format of the pre-tests is similar to that of the operational test, including the rubrics and grouping of items; but the number of items in the test can vary.

The construction of the pre-tests is much less a problem than that of finding a suitable population on which they can be piloted. Even in public examinations for schools this can give rise to difficulties. In higher education, where examinations tend to be specific to institutions rather than national, it can be even more difficult to find a pilot population which is representative of all candidates. Fortunately, while pre-testing is desirable, it is not essential. There are some hazards in using items which have not been pre-tested, but it <u>has</u> been done in school examinations and there is no reason why it should not be done, under certain circumstances, in higher education. It is useful to know the pre-tested Facility and Discrimination Indices, but with experience these can be estimated and it is by no means certain that F and D values in the pre-test will be the same as those found in

the operational test. Pre-tests can reveal ambiguities and other constructional faults in items, but so can experienced and rigorous reviewers.

If it is decided that pre-testing is not feasible, there is always the possibility of using post-test data. It is normal practice, in any event, to conduct item analysis after the use of an actual test, thus giving similar data to that obtained in pre-testing. In this way there can be accumulated a bank of items the operational characteristics of which have been recorded. (Of course, a similar bank can be built up of pre-tested items.) The item bank will be of use only if it is possible to use items more than once and, for examination purposes, this can give rise to security problems. Security, however, need not be such a problem in higher education, where the number of candidates is relatively smaller than in school examining. Tests can be more easily retrieved and accounted for. Even if complete retrieval of tests is uncertain, a bank of several hundred items has a degree of security arising from the large number of items. The chance that students could learn, or would even wish to learn, the answers to all of them is not great.

Item banking, if it is to be efficient, requires a filing and indexing system which allows rapid identification and retrieval of suitable items for test construction. The items should be classified in accordance with the following characteristics:

Subject topic
Type of ability tested
Facility index
Discrimination index.

The primary classification is usually by topic, the item itself being typed on a card with the other three characteristics being recorded in coded form on the card.

PART III

THE USE OF OBJECTIVE TESTS

1. OBJECTIVE TESTS IN EXAMINATIONS

There are two functions of objective tests in higher education, which although not completely distinct, require separate consideration. The first is their use in examinations and the second is their use as an aid to teaching and learning. This section deals with the first.

The term examination has become somewhat blurred in recent years, particularly with the advent of so-called continuous assessment. For the present purpose it is defined as a test of attainment taken at the end of a course of study. The main purpose of an examination is to distinguish between the able and less able candidates. It is necessary, therefore, that an examination should produce a spread of scores; in other words it should discriminate between candidates. Objective tests do this at least as well, and possibly better, than other forms of examination. The other advantages and disadvantages of objective tests in examinations have already been discussed, particularly with regard to their reliability and curriculum validity. This section is concerned mainly with their construction and administration.

Test Specification

The specification sets out quantitatively what the test is intended to measure. The most widely used form of specification is a two-dimensional matrix, usually called a two-way grid, which shows the proportion of items allocated to each subject area and each student attribute or ability. A notional grid of this type is shown below.

Topic / Ability	1	2	3	4	5	6	7	8	Total
A	2	2	4	2	4	2	2	2	20
B	2	2	4	8	8	4	4	8	40
C	4	8	8	2	4	4	4	6	40
Total	8	12	16	12	16	10	10	16	100%

The division of subject matter into topics, and the weighting attached to each, is a matter for decision by teachers and examiners. (These may be changed from year to year but it must then be recognized that changes do not affect the examinations alone; any change is a change in opinion about the relative importance of parts of the curriculum.) Weighting can be expediently determined by relating it to the time spent on each topic during the course but this may be an oversimplification. Some topics, particularly if they include practical work, may consume more time than their importance merits.

The division of student attributes or abilities is rather more difficult than that of subject matter. It has not been general practice in higher education to analyse

student attainment into particular skills or abilities, still less to attach specific weighting to each. In addition, there may be considerable disagreement about the form of classification of those attributes. But some thought needs to be given to this in the design of an objective test if only to avoid what is all too common - a test consisting of items which test nothing but the ability to recall previously learned information.

At school level a simple, and frequently used, classification is:

A - Items which require nothing more than the recall of information.

B - Items which require some form of processing of information in a manner with which the candidates can be expected to be familiar.

C - Items which have some aspects which are likely to be unfamiliar to the candidates, requiring the ability to apply their knowledge and to solve problems which in some important aspect they have not met previously.

Again, only teachers and examiners can decide what weight will be attached to each class of item. In higher education one would expect that items testing recall alone would have less weighting than those testing more complex skills.

In the example given above each of the cells of the matrix has been filled. Experience shows that it is almost impossible rigidly to adhere to those detailed specifications. In practice, and provided that there are no large gaps nor overweightings in the cells of the

matrix, the most one can expect is that the totals for items testing each topic and each ability in a test will approximate to the totals in the specification.

It is easy to become too complex in this aspect of design. In the early stages of the use of objective tests, the main function of the test specification is simply to point to gross discrepancies in the curriculum validity; more sophisticated matrices are rarely necessary for the purpose of assessment of attainment. (Those who have a statistical bent will note that the notional weighting allocated to various parts of the test in the specification is not necessarily the same as the actual weighting when the test operates. The latter is a function of the standard deviation of the scores and this can vary from one sub-set of part scores to another. This is rarely of concern.)

Choice and number of items

Having decided on a test specification, and having written sufficient items, the next step is to make a selection of items to match the specification. Rarely, in the early stages, is an item bank "sufficient" to allow adequate choice to meet a specification; and it is by no means unusual to find test constructors modifying their specification in order to match their items - a most improper procedure in the eyes of the test purists. But this whiff of pragmatism is not only healthy, it is a good deal more common than many test constructors care to admit to in print; and, kept within bounds, it is unlikely to make much difference. Tinkering with specifications will not give rise to any great change in the rank order of candidates, and given the

crude nature of many examinations as a measure of attainment, little if any injustice will be done. As the item bank builds up (one should think in terms of 500 items to produce annual tests of 50 or 60 items) the easier it will be to satisfy one's specification scruples.

The number of items a test can contain depends on two, sometimes opposing, factors. One is the number of items needed to match the specification, the other is the number of questions it is reasonable to ask within the time available. The time factor is usually the limiting one.

How long should a test last? At school level tests seldom exceed 90 minutes; tests are tiring and there is no reason to expect students in higher education to endure for longer periods than this. Since any item which requires processing of information, however routine in nature, should be allowed a minimum of 1½ minutes, the effective ceiling for a 90 minute test is 60 items.

Not every item will require the full 1½ minutes; items of the simple recall type, for example, need less. More than 60 items would only be included if all were of a simple recall type. Similarly, since items involving real problem solving or the analysis of complex information require correspondingly more time, the number of complex items in a test would be fewer than 60.

The lower limit on time and number of items depends on how few items can be used without losing that coverage of the curriculum which is one of the main strengths of this kind of assessment. A worthwhile objective test examination seldom has fewer than 20 items. (This does not,

of course, apply to class tests.) The effective range of numbers of items, then, is between 20 and 60 - 40 or 50 items being the norm - and the time between 30 minutes and 2 hours.

Arrangement of items

Having decided how many items to use and having chosen them to meet the test specification, the next step is to arrange them in some sort of order. Several criteria can be used for determining this.

One criterion might be to arrange the items in order of difficulty, usually difficulty increasing (i.e. F decreasing) as the test proceeds. Even a knowledge of pre-test Facility Indices does not allow this to be done with any certainty; it is barely worth the effort. An exception might be an examination intended to cover a wide ability range (the recently proposed common system of examining at 16+ would be an instance of this); but even there it would probably be better to form groups of items, each group containing items of approximately equal Facility, rather than to try to arrange them in a fine gradation of Facility from start to finish. It is common practice to put one or two easy items at the beginning of a test on the homespun psychological grounds that they will give candidates confidence and put them at their ease. This particular piece of objective test lore awaits proof; it is better to arrange the items according to some other criterion and let Facility take care of itself. In most tests the Facilities will fall in random fashion.

A more usual and more defensible practice is to arrange

items in groups corresponding to some acceptable classification of subject matter. A further refinement is to arrange the groups in accordance with a logical teaching sequence - if such a sequence is known. One of the disadvantages of objective tests, and one of the reasons why they can be so tiring for candidates, is that the candidates are required to switch very rapidly from one question to another, each question generally being unrelated to the previous one. If there is a subject-matter link within a group of questions the mental jumps involved, at least within that group, may be less demanding.

Another criterion is that items should be grouped according to type, if there is more than one type of item used within the test. In this way all multiple choice items would be grouped together, all multiple completions would be in another group, and so on. In a mixed objective test, this grouping by item type is essential; it is confusing for a student to move at random from one type of item to another. Of course subsidiary grouping can still be arranged within each item type. When only multiple choice items are used, as is most common, the problem does not arise.

At the same time as the order of arrangement of items is being considered, one should keep an eye on the sequence of keys: that is the letters which denote the correct response. It is usually thought to be undesirable to have a prolonged run of identical keys such as a long sequence of 'A's. Two or three in a sequence do not matter, but if a student finds that he is giving the same key in a longer sequence, he may begin to think that

at least some of his answers are incorrect, particularly if he knows that it is usual to have a greater frequency of change. Generally it is not necessary to go to great trouble to randomize the sequence of keys; it is remarkable how this is often found to have taken care of itself when the test has been completed. But the sequence should be checked, and, if necessary, altered to avoid long runs of the same key.

Similarly, a check should be made of the total number of times each key occurs in a test. The incidence of each should be approximately the same. For example in a 40 item test, consisting of items each with 4 responses, then A,B,C, and D, should individually occur as the correct response about 10 times. Again, this need not be applied too rigidly; anything between 8 and 12 times will do. Outside those limits it might be advisable to alter the sequence of responses in order to avoid overweighting any one of them. Occasionally one comes across a random guesser who opts for 40 'A's or 'B's - whichever his fancy lights on. It is just possible that luck, or lack of it, in such guesswork, could be exaggerated by an unequal distribution of keys but this is not a sufficient reason for equalizing numbers of keys. A more important point to consider is the student who, knowing that it is usual to equalize the numbers, finds that he has chosen an abnormally high or low number of a particular key. He may then be tempted to change the answer for purely statistical reasons. This problem is similar to that of sequence of keys.

Lay-out of the test

Clarity of lay-out is important. It is easy enough for

a student to become confused when dealing with a single complex item: it is doubly so if the layout of the items is not easily followed. The temptation to save space and paper should be resisted. It is possible to arrange items in a double column on each side of the paper, but only if the items themselves are short and simple. Even then, the order and numbering of the items should be logical and consistent. This is particularly so if the keys have to be recorded on a separate answer sheet and not directly on the item itself.

An item should be complete within one side of paper. If it is started on one side and carried over to another, the candidate may have to refer back and this can be distracting and time-consuming especially if the two parts are not on facing pages. The same applies to the kind of complex rubric used with multiple completion items - such rubrics should be repeated on each page in summary form so that candidates do not have to turn a page to remind themselves of the form in which they must answer.

General Rubric

The rubric at the beginning of a test should give the following information.

1. How long the test will last.
2. How many marks are given for each correct response. (Almost invariably one mark.)
3. How incorrect responses will be scored (usually zero, but sometimes a penalty mark is awarded).
4. That there is only one correct response to each item (this avoids the need to write 'one of' in the stem of every item).

5. What equipment (pencils, erasers, etc.) will be required.
6. How to record the chosen response.
7. How to correct a chosen response.
8. Some guidance on the sequence of attempts; for example:

   "You are advised to start at the first item and continue in sequence until the last. If at the first attempt you cannot decide on an answer to an item, do not spend too much time on it; continue to the end then, if you have time to spare, return to those items you have not answered."
9. It is not usual to advise candidates, in writing, to guess when they are not sure of the answer. Before a non-penalty test, however, a teacher would have to have an incredibly tender moral sense not to advise his students verbally to do so. In practice random guessing is rare; students may say that they have guessed, but actually they mean that they were not sure whether they were right, so they chose the most likely - a worthy reason for any choice.

Responding

There are various ways in which responses can be recorded. In large scale examining the use of some form of answer sheet is preferred; that is a sheet which is either separate or capable of being detached from the test itself. There are three main types of answer sheet, outlined briefly below, but in all three the use of a soft lead pencil is advocated because it facilitates correction if a

candidate should have a change of mind.

The first type of answer sheet requires the candidate to write the letter or number corresponding to the chosen response to each item. For example:

BLANK ANSWER SHEET	COMPLETED ANSWER SHEET
1. ☐ 2. ☐ 3. ☐	1. [A] 2. [C] 3. [A]
4. ☐ 5. ☐ 6. ☐	4. [D] 5. [D] 6. [B]
7. ☐ 8. ☐ etc.	7. [A] 8. [B] etc.

In the second type of answer sheet the candidates must mark, usually by a circle, or a cross or a tick, the letter or number corresponding to the chosen response; these letters or numbers being already printed on the answer sheet. For example:

BLANK ANSWER SHEET	COMPLETED ANSWER SHEET
1. A B C D	1. A B C D
2. A B C D	2. A B C D
3. A B C D	3. A B C D
4. A B C D etc.	4. A B C D

There are, of course, variants of both of these methods of answering, but the principle remains.

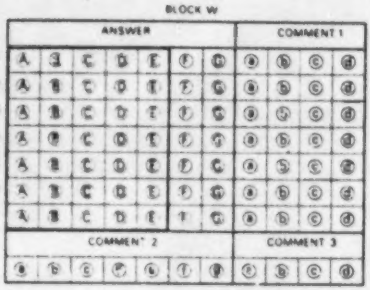
The third type is used mainly when tests are marked by machine. It requires candidates to fill in completely a small lettered or numbered space by means of a pencil mark. When fed through a scoring machine, which can detect the positions of the pencil marks and which is programmed to distinguish between correct and incorrect scores, each

sheet is automatically scored. For example:

BLANK ANSWER SHEET				COMPLETED ANSWER SHEET			
1. A===	B===	C===	D===	1. A▬▬	B===	C===	D===
2. A===	B===	C===	D===	2. A===	B▬▬	C===	D===
3. A===	B===	C===	D===	3. A===	B===	C===	D▬▬
4. A===	B===	C===	D===	4. A===	B===	C▬▬	D===

etc.

Another form of this kind of answer sheet is a block such as the Open University uses for computer scoring. For example:



(Comment 1 cells are to indicate reasons for not answering questions.

Comment 2 cells are to indicate the degree of difficulty of the block.

Comment 3 cells are to show how well students think the block complements teaching.)

The Open University teaches partly by correspondence; the Comment cells are to provide feedback to teaching staff.

In examinations with small numbers of candidates the chosen response can be indicated by making a mark directly onto the item itself. This avoids the possibility of error of translation on the part of the student when his eyes move from the item to a separate sheet. However, a test which has had marks made on it cannot be used again. In general, a separate answer sheet of some sort is advocated.

2. OBJECTIVE TESTS IN TEACHING AND LEARNING

In examinations objective tests are used deliberately to give rise to a spread of scores and hence form a basis for grading candidates. When they are used as a teaching or learning aid, however, this discriminatory function, although inevitably it will arise, becomes secondary. The main functions in teaching and learning are:

* to provide evidence of attainment,
* to provide evidence of the effectiveness of the teaching and learning,
* to assist in the identification of weakness in teaching and learning.

In short, the objective tests are used for diagnostic purposes; they are a guide to teachers and students about what has been achieved, what remedies need to be taken, and what should be the next steps in the teaching and learning process.

Clearly this use of objective tests is more suitable for some kinds of learning than others. Learning which is open-ended and discursive, where there are no clearly right or wrong answers, is less likely to be assessed by a mode which essentially looks for a fixed response and conformity with the teacher's preconceived judgment about what constitutes the best answer. But those concerned with the more open-ended type of learning should not dismiss objective tests out of hand. In a teaching situation it may well be that the choice by a student of a response which was not that chosen by the teacher could give rise to very fruitful discussion about the reasons for the choice.

To give an example.[1]

> "Given that the price elasticity of demand for the products concerned is zero, a successful pay claim in an industry will have to be paid for, other things being equal, by the
>
> A shareholders, as profits will be reduced
> B workers, as all those whose marginal revenue product is below the new pay level will no longer be employed
> C consumers of the products
> D the general public, since prices will have risen
> E shareholders and consumers in some combination, which will depend on the latter's ability to find substitutes".

The key to this item is C; but without doubt some students would answer otherwise because the variables in such a situation are complex and so may be the links between cause and effect. In a teaching situation, in which discussion between student and teacher is possible after the test, students who had chosen A or E, for example, could be asked for reasons for their choice. The resulting discussion could reveal some interesting conceptions or misconceptions by both teacher and student; and what had started as essentially a fixed response, convergent situation could open out into a free response, divergent situation. If a majority of students had chosen responses other than C, the discussion would be even more lively.

---

(1) This example is item 23 in Specimen Paper (1975) G.C.E. Economics A Level, published by The Associated Examining Board, Aldershot.

This brings us to a fundamental point with regard to the use of objective tests as a teaching aid: the favoured WRONG responses can be more important and revealing to teachers and students than the correct responses. In any analysis of objective test items used for diagnostic purposes it is essential that the favoured wrong response be identified.

If such diagnosis is to be effective, feedback of item analysis to both teacher and students must be as rapid as possible. Identification of error, and discussion of causes of the error, needs to be done before the event has grown cold. It follows that self-assessment by the students as soon as the test has been completed is one way, and possibly the best way, of achieving this. To do so, the students must be given the keys to the test; they must also be asked to identify the reasons for their incorrect responses and, if necessary, be asked to report on those reasons.

Self-assessment necessarily imposes a limitation on the function of objective tests. It is too idealistic to suppose that students will conscientiously avoid reference to the keys if they know that the test is to be used as part of a continuous system of assessment affecting their final grade. It is tempting for teachers to try to have the best of both worlds by using the test for both diagnostic and discriminatory purposes. Far better to go firmly for one or the other. Once students are convinced that the sole purpose of the tests is to diagnose their difficulties and guide their subsequent study, self-assessment becomes not only feasible but desirable.

The lower limit on the number of items in a test suggested for examination purposes (20) need not apply to diagnostic tests. 20 items may be needed in order to get respectable test statistics in an examination, but those statistics matter less in diagnostic tests. The number of items in a test will depend largely on how frequently you test and hence on the amount of subject matter to be tested. 10 items could constitute a reasonable test; use more if necessary, but fewer would make testing barely worth while.

Norm Referenced and Criterion Referenced Tests

The analysis of items for diagnostic tests is less strict in the limits it imposes than is the case with discriminatory tests. The facility index for a discriminatory test ideally will fall within a narrow band of values centred on 50% for each item, while much higher facilities are acceptable both for each item in a diagnostic test and for the test as a whole. When properly designed and executed teaching is intended to lead to successful rather than unsuccessful learning. Diagnostic tests, properly related to teaching design - that is tests with high curriculum validity - should therefore yield high not low scores. If diagnostic test scores are low then something is wrong with either teaching or test or both; one or both should be changed.

The discrimination index is hardly relevant to diagnostic tests at all. The criterion of a good diagnostic test is not how well it separates the good student from the bad, but how well it identifies what has been learned and what has not been learned. In other words a diagnostic test is not norm referenced. Whereas in a discriminatory test,

when discrimination is taken to the extreme, the substance of the test matters less than the discrimination index, in a diagnostic test the substance is all important. It follows, of course, that what is expected to be learned must be explicitly defined in terms of both subject matter and student skills, and items should be designed to match that learning. These explicit outcomes become the criteria for judging whether learning has taken place or not; the test is said to be criterion referenced.

It is necessary to say at this point that the distinction between norm referenced tests and criterion referenced tests is not as polarized as the foregoing paragraph may suggest. Maximum discrimination is never the sole basis for constructing an objective test even for strongly graded examinations; if it were, the subject matter need be only trivial or even irrelevant in the curriculum context. Discriminatory tests <u>are</u> constructed according to a specification of subject matter and skills and in that sense they are criterion referenced; and they can be used to identify problems in teaching and learning related to the whole examination population.[1] But they are not designed to identify the learning problems of individuals. Two candidates could have identical scores of, say, 25 out of 50; but the 25 items each answered correctly could be completely different. <u>Statistically</u> they have performed equally. In terms of <u>learning</u> they have learned different things and exhibited different skills.

Conversely, it is too much to expect that a diagnostic

---

(1) MATHEWS, J.C. & LEECE, J.R. (1976) Schools Council Examinations Bulletin 33. *Examinations: their use in curriculum evaluation and development.* London. Evans/Methuen.

test, although strongly criterion referenced, will not discriminate. In any group of students, however close their abilities, some inevitably will do better than others. Therein lies the temptation for teachers to use the scores as part of the final assessment.

Individualized Learning

Individualized learning is having something of a vogue. It ranges from highly complex computer assisted learning systems, through various kinds of programmed learning, to some very simple written materials which are little more than annotated texts. Those who are inclined to react against the fashionable trends in education may point to books as the original, and still the most important, self-instructional material. But individualized learning materials (loosely defined as materials or learning systems which do not require the presence of a teacher) have at least one thing which makes them different from books: they have within themselves assessment systems which enable a student to measure his progress through the material. Tests, often of the objective type, are built into the self-instructional materials. The student who fails them at a first attempt should retake them before progressing. In some systems he may, however, be led to remedial learning.

Space is too limited to describe the highly sophisticated, meticulously validated, and scrupulously edited, tests which have been devised in large scale and extravagantly financed self-instructional projects. Anyway, the purpose of this booklet is to assist the teacher in higher education working alone or with a small group of colleagues. Nevertheless, there are some areas in

higher education in which teachers are designing materials from which the students can learn in the teacher's absence. Sometimes these materials are essentially remedial for those who have not grasped (or possibly failed to attend) a particular lecture or seminar. Sometimes they may be concerned with a specific skill or body of knowledge which is better learned at a pace suited to individual students and not in a large group teaching situation. Tests are an essential adjunct to any such material. They have three functions: pre-tests to decide whether a student is ready to study the material, concurrent tests to enable the student to estimate his progress through the material, and terminal tests to measure his final attainment.

These tests do not have to be of the objective type, particularly if the learning is open-ended. But when learning is designed to gain mastery of, say, a clearly defined skill, an objective test, precise in both construction and marking, seems to be the best possible means of measuring whether or not mastery has been achieved.

Self-instructional systems depend greatly for their effectiveness on the choice of the most appropriate size for the sub-units of learning which make up the whole. Skill in designing these steps, which in sequence make up the self-instructional package, is linked to the design of tests of attainment. Both require the clear specification of the outcomes of learning. However, designing learning units is a specialized field beyond the scope of this booklet; here the point to be made is that each step in a self-instructional system is preceded and terminated by a test. The first assesses whether a student is

ready to start, the second assesses what he has learned and what remedial learning, if any, is now necessary. A network of diagnostic tests which charts each student's progress through the material is thus systematically built up.

PART IV

IN CONCLUSION

It has often been said that one of the best ways of coming to understand the essential meanings of a subject is to try to teach it. In the author's experience, even greater insights can accrue from the attempt to construct objective tests in a subject. Nothing is more calculated to force one to analyse the meanings of statements, and questions, and answers, than to try to put them into one of the forms of objective items. Criteria for judgment, previously assumed, have to be made explicit and defended in front of one's colleagues; information, previously taken to be true, has to be qualified to avoid misinterpretation; reasons, glibly given in a lecture, look poor things in the cold, written form of a relationship analysis item. There is no better training for teachers, whether they be subject specialists or not. So, even if you have doubts about them, at least try to write some - the exercise might be good for you if not for your students.

APPENDIX - EXAMPLES OF OBJECTIVE ITEMS

(The sources of these items are given at the end of this appendix)

MULTIPLE CHOICE

Item 1

Which one of the following words or phrases describes the process of engaging National Government officials in Community decision-making?

A Decalogue
B Engrenage
C Conseil d'etat
D Palazzi
E Détournement de pouvoir

Item 2

'Whenever the factors of production are combined to produce any given commodity, the cost to society is to be reckoned in terms of the alternative goods which those factors could have produced.' This statement seeks to explain the concept of

A average cost
B factor cost
C opportunity cost
D comparative cost
E marginal cost

Item 3

Chlorine trifluoride and boron trifluoride have molecular formulae of the type XY_3, and their molecular shapes are

A planar (both)
B planar (ClF_3) and pyramidal (BF_3)
C pyramidal (ClF_3) and planar (BF_3)
D pyramidal (both)

Item 4

Of the following pairs of organs, those characteristically affected in glycogen storage disease are

A the heart and the liver
B the pancreas and the kidney
C the liver and the brain
D the liver and the pancreas
E the heart and the pancreas

Items 5-8

A sixty-year-old woman suddenly developed fever, a productive cough, and pain the right side of the chest which was accentuated by breathing. She is admitted to the hospital four days after the onset of her symptoms. Her body temperature is 102.0°F (38.9°C). Examination of the chest reveals dullness, absent tactile fremitus, and markedly diminished breath sounds over the lower half of the right lung. The leukocyte count is 24,000 per cu mm. The sputum is pink in color and tenacious.

5. The predominant organism cultured from the sputum is most likely to be

   A pneumococcus
   B *Klebsiella*
   C tubercle bacillus
   D beta-hemolytic streptococcus
   E staphylococcus

6. Roentgenogram of the chest will show evidence of

   A consolidation of the right lower lobe
   B consolidation of the right lower lobe and pleural fluid
   C massive atelectasis of the right lung
   D a large cavity in the right lower lobe
   E pulmonary infarction

7. The patient fails to respond to penicillin therapy. The next procedure should be

   A administration of a broad spectrum antibiotic
   B to repeat a culture of the sputum
   C bronchograms
   D needle aspiration of the pleural cavity
   E bronchoscopy

8. Further diagnostic study discloses

   A bronchiectasis
   B tuberculosis
   C lung abscess
   D empyema
   E bronchogenic carcinoma

CLASSIFICATION SETS

Items 9-13

DIRECTIONS: Each group of questions below consists of five lettered headings followed by a list of numbered words or phrases. For each numbered word or phrase select the ONE heading that is most closely related to it and blacken the corresponding space on the answer sheet.

A Hypertrophy of the left ventricle
B Cor pulmonale
C Mitral and aortic stenosis
D Subpulmonic stenosis
E Congestive failure without cardiac enlargement

9. Long-standing silicosis
10. Constrictive pericarditis
11. Rheumatic heart disease
12. Systemic hypertension
13. Congenital anomaly

Items 14-20

Match each of the 7 names listed below with one of the organizations or events (A-G) which is most closely associated with them. The numbers (U1-U7) in the first list correspond to the row numbers of the BLOCK. The letters (A-G) in the second list correspond to the column letters of the same BLOCK. For example, if you think Reginald Maudling (U1) is most closely associated with the Messina Conference (G), pencil out G in row U1.

U1 Reginald Maudling
U2 George Marshall
U3 Georges Pompidou
U4 Paul-Henri Spaak
U5 General de Gaulle
U6 René Pleven
U7 Robert Schuman

A European Defence Community
B Luxembourg Compromise
C Organization of European Economic Cooperation
D European Free Trade Area
E European Coal and Steel Community
F Hague Summit
G Messina Conference

BLOCK U

	ANSWER							COMMENT 1			
1	A	B	C	D	E	F	G	0	1	2	3
2	A	B	C	D	E	F	G	0	1	2	3
3	A	B	C	D	E	F	G	0	1	2	3
4	A	B	C	D	E	F	G	0	1	2	3
5	A	B	C	D	E	F	G	0	1	2	3
6	A	B	C	D	E	F	G	0	1	2	3
7	A	B	C	D	E	F	G	0	1	2	3
	COMMENT 2							COMMENT 3			
	0	1	2	3	4	5	6	0	1	2	3

(Note that this block is used for computer marking and the additional rows and columns are to provide tutors with feedback. See the notes on page 42.)

MULTIPLE COMPLETION

Items 21-23

DIRECTIONS: For each of the incomplete statements below, ONE or MORE of the completions given is correct. On the answer sheet blacken the space under

A if only *1,2, and 3* are correct,
B if only *1 and 3* are correct,
C if only *2 and 4* are correct,
D if only *4* is correct,
E if all are correct.

Directions Summarized

A	B	C	D	E
"	"	"	"	"
"	"	"	"	"
"	"	"	"	"
1,2,3 only	1,3 only	2,4 only	4 only	all are correct

21. Portions of the humerus covered with periosteum include the

(1) greater tubercle
(2) lesser tubercle
(3) anatomical neck
(4) surgical neck

22. The walls of the true pelvis are formed in part by the

(1) piriformis muscles
(2) pubis
(3) obturator internus muscles
(4) sacrum

23. The antimicrobial activity of the sulfonamides is

(1) decreased in the presence of marked suppuration
(2) prolonged by the concomitant use of probenecid
(3) antagonized by local anesthetic agents such as procaine
(4) limited to gram-negative organisms

Items 24-26

Directions

For questions 24 to 26 ONE or MORE of the responses is (are) correct. Decide which of the responses is (are) correct and then choose

A if only (i), (ii) and (iii) are correct
B if only (i) and (iii) are correct
C if only (ii) and (iv) are correct
D if only (iv) is correct

DIRECTIONS SUMMARISED			
A (i) (ii) (iii) only	B (i) (iii) only	C (ii) (iv) only	D (iv) only

24. In the chemistry of Group VII

(i) all the halogens form oxyacids of the types HXO_3 and HXO_4
(ii) hypohalous acids HOX are formed by all the halogens except fluorine
(iii) fluorine forms only one oxide, F_2O
(iv) F_2O_5 is a thermally stable white solid

25. As a result of undergoing a d,α nuclear reaction, the isotope $^{114}_{48}Cd$ would become

(i) of lower nuclear mass
(ii) of lower atomic number
(iii) $^{112}_{47}Ag$
(iv) $^{116}_{46}Pd$

26. Which of the following metals in the given oxidation state have partially-filled d orbitals?

(i) Sc(III)
(ii) Ti(III)
(iii) Zn(II)
(iv) Cu(II)

Items 27-29

Directions. For each of the questions below, ONE or MORE of the responses given is (are) correct. Decide which of the responses is (are) correct. Then choose

A if 1,2 and 3 are correct
B if 1 and 2 only are correct
C if 2 and 3 only are correct
D if 1 only is correct
E if 3 only is correct

Directions Summarised				
A	B	C	D	E
1,2,3 correct	1,2 only	2,3 only	1 only	3 only

27. *Estimated Population of the U.K.*

Year	*Millions*
1980	60
2000	72

(Based on the assumptions that death rates decline, annual births increase and net migration is nil.)

Which of the statements below is (are) reasonable deductions from the above information?

1 Housing demand will increase
2 Unemployment will increase
3 Income per head will decrease

28. Which of the following advantages enjoyed by large-scale industry is (are) external economies of scale?

1 Discounts from bulk buying
2 Growth of ancillary industries
3 Development of specialist markets

29. Which of the following would probably result in a shift in the demand curve for cameras?

1 A decrease in the price of cameras
2 An increase in the tax on cameras
3 A decrease in the price of film

Items 30-41

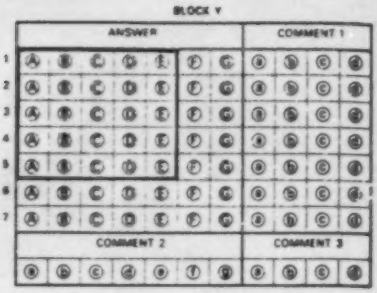
BLOCK Y

This block carries 25 per cent of the total marks for this assignment.

In the matrix below are 25 statements about the Commission. Which twelve statements would you accept concerning the Commission's competence and functions?

Pencil out the appropriate cells in BLOCK Y using rows 1-5 and letters A-E only. Each correct answer will count towards your score. You will be penalized for penciling out more than 12 cells.

(See note accompanying Items 14-20 above.)



	A	B	C	D	E
1	Its position requires non-political administrative decision-taking	It negotiates for the Community in trade talks	It has the power to insist on adoption of its proposals in all cases	Its headquarters are in Luxembourg	A major function is to see that the Treaties are respected and applied
2	The Presidency changes every year	Among its formal powers is responsibility for initiating policy proposals	It may take Member States before the European Court	It is remote from officials of Member States	The Luxembourg compromise extended its freedom of action
3	It has regularly used the tactic of open confrontation with the members to make progress in the Communities	It shares many common interests with the European Parliament	Its staff is larger than that of the British Government	It involves members of National Governments in decision-making	It favours the grouping of interest groups on a Community scale
4	It is the executive of the EEC, ECSC and Euratom	It may not impose fines	It implements Community policy	It is responsible to the European Parliament	It is a collegiate body
5	Its powers of action are independent of both Council of Ministers and Member States	In 1965 the French withdrew all their officials from the Commission	Commissioners are appointed by the National Governments	Its political role and its mediatory role are inherently contradictory	It has 18 members

n from 5 MULTIPLE COMPLETION

Items 42-43

ONE or MORE of the answers to these questions may be correct.
(This type of item is normally marked by computer.)

42. Mark the appropriate box for each answer which you judge to be TRUE (CORRECT).

For a non-competitive inhibitor, the degree of inhibition depends upon:

- ☐ a) the concentration of the inhibitor
- ☐ b) the concentration of the substrate
- ☐ c) the K_i for the inhibitor
- ☐ d) the K_m for the substrate
- ☐ e) the turnover number of the enzyme.

43. ALL FIVE of the alternative answers require a response. For those choices judged to be TRUE (CORRECT) mark the box in the T column. For those choices judged to be FALSE (INCORRECT) mark the box in the F column.

T	F	Myoglobin:
☐	☐	a) contains a high proportion of α-helix
☐	☐	b) is monomeric
☐	☐	c) consists of a number of subunits
☐	☐	d) has a sigmoid oxygen-binding curve
☐	☐	e) has a hyperbolic oxygen-binding curve.

RELATIONSHIP ANALYSIS

Items 44-45

DIRECTIONS: Each question below consists of an *assertion* (statement) in the left-hand column and a *reason* in the right-hand column. On the appropriate line of the answer sheet blacken the space under

A if both assertion and reason are true statements and the reason is a *correct explanation* of the assertion;

B if both assertion and reason are true statements but the reason is *NOT a correct explanation* of the assertion;

C if the assertion is true but the reason is a false statement;

D if the assertion is false but the reason is a true statement;

E if both assertion and reason are false statements.

Directions Summarized			
A -	True	True	Reason is a correct explanation
B -	True	True	Reason is NOT a correct explanation
C -	True	False	
D -	False	True	
E -	False	False	

	Assertion		*Reason*
44.	The blood sugar level falls precipitously after hepatectomy	BECAUSE	the glycogen of the liver is the principal source of blood sugar.
45.	Anatomic diagnosis of an adenoma of the islet cells of the pancreas means that the patient has hyperinsulinism	BECAUSE	all islet cell adenomas are functional and secrete insulin.

Items 46-49

Directions summarized:

	Assertion	Reason	
A	True	True	Reason is a correct explanation.
B	True	True	Reason is *NOT* a correct explanation.
C	True	False	
D	False	True	
E	False	False	

Remember that your answer should be based on the views contained in Unit 1 and *NOT* on your personal views.

	Assertion	Reason
46.	Before the Second World War, Britain consistently opposed schemes for European political unity	Britain had the most widespread overseas interests of the European states.
47.	The post-war Labour government favoured European federation	It sought unimpeded control of the national economy.
48.	Although opposed to the surrender of national sovereignty, British governments in the 1950s supported moves towards functional integration	They regarded such integration as strengthening the economy, without requiring the surrender of political authority.
49.	In the early 1960s, de Gaulle feared that British entry into the Community would introduce an American Trojan Horse	The election of John F. Kennedy to the American presidency strengthened Britain's role as 'diplomatic tutor'.

Items 50-53

DIRECTIONS SUMMARISED			
	Assertion	Reason	
A	True	True	Reason is a correct explanation
B	True	True	Reason is NOT a correct explanation
C	True	False	
D	False	True	

	Assertion	Reason
50.	(i) For light atoms, nuclear fusion is energetically feasible, while for heavy atoms nuclear fission is energetically feasible	(ii) The plot of binding energy per nucleon against atomic weight has a minimum atomic weight 65
51.	(i) Chloroform has a dipole moment	(ii) The chloroform molecule is tetrahedral
52.	(i) Phosphorus forms no molecular oxides corresponding to NO, NO_2 and N_2O	(ii) Phosphorus unlike nitrogen, possesses energetically accessible d orbitals
53.	(i) Phosphorous acid (H_3PO_3) is a dibasic acid	(ii) The energetically favoured structure for this compound involves a phosphorus-oxygen double bond

Items 54-57

	Directions Summarised		
	First Statement	Second Statement	
A	True	True	2nd statement is a correct explanation of, or reason for, the first
B	True	True	2nd statement is NOT a correct explanation of, or reason for, the first
C	True	False	
D	False	True	
E	False	False	

	FIRST STATEMENT	SECOND STATEMENT
54.	Increased government spending increases inflation under all conditions	Government spending is not offset by any form of production
55.	Private monopolies usually charge prices at the lowest possible point	Monopolies may enjoy some economies of scale
56.	One of the main aims of the Trades Unions is to increase the real wages of their members	In order to maximise the real wages of their members in a capitalist society Trades Unions aim at all firms making zero profits
57.	Exports are a 'leakage' from the economy and should be deducted from the National Income	Exports are sent overseas and and so are not available for internal use

SOURCES OF THE ITEMS IN THIS APPENDIX

(1) Items 1, 14-20, 30-41 and 46-49 are taken from:

Computer Marked Assignment P933 41 (1974)
Supplementary Material P933 1-8 for a Post-experience course *P933 European Economic Community*

Milton Keynes. The Open University.

(2) Items 2, 27-29 and 54-57 are taken from:

The Associated Examining Board Specimen Paper (1975)
General Certificate in Education
Economics - Advanced Level

Aldershot. The Associated Examining Board.

(3) Items 3, 24-26 and 50-53 are taken from:

MATHEWS, J.C. (undated) *Objective Testing - A Workshop Report*

London. The Chemical Society: Education Division.

(4) Items 4, 5-8, 9-13, 21-23 and 44-45 are taken from:

HUBBARD, J.P. and CLEMANS, W.V. (1961) *Multiple Choice Examinations in Medicine*

Philadelphia. Lea & Febiger.

(5) Items 42-43 are adapted from:

MORGAN, M.R.J. (personal communication) *Multiple Choice Examination for Biochemistry 201: explanatory handout*

University of Lancaster. Department of Biological Sciences.

TEACHING IN HIGHER EDUCATION SERIES

1. Lecturecraft
2. Why Lecture?
3. Working Together
4. Participatory Discussion
5. A Kind of Learning
6. Patterns and Procedures
7. Inside Assessment
8. Essays
9. The Use of Objective Tests

School of Education, University of Lancaster